IEEE Control Systems Letters paper presented at
2023 62nd IEEE Conference on Decision and Control (CDC)
December 13-15, 2023. Marina Bay Sands, Singapore

Revisiting LQR Control from the Perspective of Receding-Horizon Policy Gradient

Xiangyuan Zhang Tamer Başar

Abstract— **We revisit in this paper the discrete-time linear quadratic regulator (LQR) problem from the perspective of receding-horizon policy gradient (RHPG), a newly developed model-free learning framework for control applications. We provide a fine-grained sample complexity analysis for RHPG to learn a control policy that is both stabilizing and ϵ-close to the optimal LQR solution, and our algorithm does not require knowing a stabilizing control policy for initialization. Combined with the recent application of RHPG in learning the Kalman filter, we demonstrate the general applicability of RHPG in linear control and estimation with streamlined analyses.**

I. Introduction

Model-free policy gradient (PG) methods promise a universal end-to-end framework for controller designs. By utilizing input-output data of a black-box simulator, PG methods directly search the prescribed policy space until convergence, agnostic to system models, objective function, and design criteria/constraints. The general applicability of PG methods leads to countless empirical successes in continuous control, but the theoretical understanding of these PG methods is still in its early stage. Stemmed from the convergence theory of PG methods for general reinforcement learning tasks [1], [2], a recent thrust of research has specialized the analysis for the convergence and sample complexity of PG methods into several linear state-feedback control benchmarks [3]–[10]. However, incorporating imperfect-state measurements leads to a deficit of most, if not all, favorable landscape properties crucial for PG methods to converge (globally) in the state-feedback settings [10], [11]. Even worse, the control designer now faces several challenges unique to control applications: a) convergence might be toward a suboptimal stationary point without system-theoretic interpretations; b) provable stability and robustness guarantee could be lacking; c) convergence depends heavily on the initialization (e.g., the initial policy should be stabilizing), which would be challenging to handcraft; and d) algorithm could be computationally inefficient. These bottlenecks blur the applicability of model-free PG methods in real-world control scenarios since the price for each of the above disadvantages could be unaffordable.

Classic theories provide both elegant analytic solutions and efficient computational means (e.g., Riccati recursions) to a wide range of control problems [12]–[14]. They further reveal the intricate structure in various control settings and offer system-theoretical interpretations and guarantees to their characterized solutions. They suggest that, compared to viewing the dynamical system as a black box and studying the properties of PG methods from a (nonconvex) optimization perspective, it is better to incorporate those properties unique to decision and control into the design of learning algorithms.

In this work, we revisit the classical linear quadratic regulator (LQR) problem from the perspective of the newly-developed receding-horizon PG (RHPG) framework [15], which fuses Bellman's principle of optimality into the development of a model-free PG framework. First, RHPG approximates infinite-horizon LQR using a finite-horizon problem formulation and further decomposes the finite-horizon problem into a sequence of one-step sub-problems. Second, RHPG solves each sub-problem recursively using model-free PG methods. To accommodate the inevitable computational errors in solving these sub-problems, we establish the generalized principle of optimality that bounds the accumulated bias by controlling the inaccuracies in solving each sub-problem. We characterize the convergence and sample complexity of RHPG in Section III-D and emphasize that the RHPG algorithm does not require knowing a stabilizing initial control policy *a priori*.

Compared with [3], [5], we provide a new parametrization and perspective in learning for control with performance guarantees and streamlined analyses. By unifying existing control theory into the design of learning algorithms, our work and [15] aim to explore a promising path toward the theoretical foundation of PG methods in addressing partially observable and nonlinear control through the lens of RHPG. Compared with [9], [16], [17] that have also removed the assumption on an initial stabilizing point in LQR, we provide more explicit choices of various parameters in the RHPG algorithm. Moreover, the RHPG framework can be directly extended to solve output-feedback control problems [15]. At the same time, the results for discounted LQR can hardly be generalized further since its optimization landscapes are identical to those in un-discounted LQR, with essentially scaled versions of system parameters. We defer an extensive literature review to the technical report [18].

II. Preliminaries

A. Infinite-Horizon LQR

Consider the discrete-time linear dynamical system[1]

$$x_{t+1} = Ax_t + Bu_t, \tag{2.1}$$

XZ and TB are with the Department of ECE and CSL, the University of Illinois Urbana-Champaign, Urbana, IL 61801 USA (e-mails: {xz7, basar1}@illinois.edu). Research of the authors was supported in part by ARO MURI Grant AG285 and in part by AFOSR Grant FA9550-19-1-0353.

[1]For extensions to stochastic LQR with i.i.d. additive noises, as well as the setting with an arbitrary (deterministic) initial state, see Sec. H of [18].

where $x_t \in \mathbb{R}^n$ is the state; $u_t \in \mathbb{R}^m$ is the control input; $A \in \mathbb{R}^{n\times n}$ and $B \in \mathbb{R}^{n\times m}$ are system matrices unknown to the control designer; and the initial state $x_0 \in \mathbb{R}^n$ is sampled from a zero-mean distribution $\mathcal{D}$ that satisfies $\mathrm{Cov}(x_0) = \Sigma_0 > 0$. The goal in the LQR problem is to obtain the optimal controller $u_t = \phi_t(x_t)$ that minimizes the cost

$$\mathcal{J}_\infty := \mathbb{E}_{x_0\sim\mathcal{D}}\left[\sum_{t=0}^{\infty}\left(x_t^\top Q x_t + u_t^\top R u_t\right)\right], \tag{2.2}$$

where $Q > 0$ and $R > 0$ are symmetric positive-definite (pd) weightings chosen by the control designer. For the LQR problem as posed to admit a solution, we require (A, B) to be stabilizable. Note that here $Q > 0$ implies the observability of $(A, Q^{1/2})$. Then, the unique optimal LQR controller is linear state-feedback, i.e., $u_t^* = -K^* x_t$, and $K^* \in \mathbb{R}^{m\times n}$, which with a slight abuse of terminology we will call optimal control policy, can be computed by

$$K^* = (R + B^\top P^* B)^{-1} B^\top P^* A, \tag{2.3}$$

where P^* is the unique positive definite (pd) solution to the algebraic Riccati equation (ARE)

$$P = Q + A^\top PA - A^\top PB(R + B^\top PB)^{-1}B^\top PA. \tag{2.4}$$

Moreover, the optimal control policy K^* is guaranteed to be stabilizing, i.e., $\rho(A - BK^*) < 1$. Therefore, we can parametrize LQR as an optimization problem over the policy space $\mathbb{R}^{m\times n}$, subject to the stability condition [3]:

$$\min_{K}\ \mathcal{J}_\infty(K) = \mathbb{E}_{x_0\sim\mathcal{D}}\left[\sum_{t=0}^{\infty}\left(x_t^\top (Q + K^\top R K) x_t\right)\right] \tag{2.5}$$

$$\text{s.t.}\quad K \in \mathcal{K} := \{K \mid \rho(A - BK) < 1\}. \tag{2.6}$$

Theoretical properties of model-free (zeroth-order) PG methods in solving (2.5) have been well understood [3], [5], [19]. In particular, the objective function (2.5), even though nonconvex, is coercive and (globally) gradient dominated [19]. Hence, if an initial control policy $K_0 \in \mathcal{K}$ is known *a priori*, then any descent direction of the objective value (e.g., vanilla PG) suffices to ensure that all the iterates will remain in the interior of $\mathcal{K}$ while quickly converging toward the unique stationary point. Removing the assumption on K_0 (that an initial stabilizing policy can readily be found) has remained an active research topic [9], [16], [17].

B. Finite-Horizon LQR

The finite-N-horizon version of the LQR problem is also described by the system dynamics (2.1), but with the objective function summing up only up to time $t = N$. Similar to (2.5), we can parametrize the finite-horizon LQR problem as $\min_{\{K_t\}} \mathcal{J}\big(\{K_t\}\big)$, where

$$\mathcal{J}\big(\{K_t\}\big) := \mathbb{E}_{x_0\sim\mathcal{D}}\left[\sum_{t=0}^{N-1} x_t^\top (Q + K_t^\top R K_t) x_t + x_N^\top Q_N x_N\right], \tag{2.7}$$

and Q_N is a symmetric pd terminal-state weighting to be chosen. The unique optimal control policy in the finite-horizon LQR is time-varying and can be computed by

$$K_t^* = (R + B^\top P_{t+1}^* B)^{-1} B^\top P_{t+1}^* A, \tag{2.8}$$

where P_t^*, for all $t \in \{0, \cdots, N-1\}$, are generated by the Riccati difference equation (RDE) starting with $P_N^* = Q_N$:

$$\begin{aligned} P_t^* &= Q + A^\top P_{t+1}^* A \\ &\quad - A^\top P_{t+1}^* B(R + B^\top P_{t+1}^* B)^{-1} B^\top P_{t+1}^* A. \end{aligned} \tag{2.9}$$

Theoretical properties of zeroth-order PG methods in addressing (2.7) have been studied in [7], [8]. Compared to the infinite-horizon setting (2.5), the finite-horizon LQR problem (2.7) is also a nonconvex and gradient-dominated problem, but it does not naturally require the stability condition (2.6).

III. Receding-Horizon Policy Gradient

A. LQR with Dynamic Programming

It is well known that the solution of the RDE (2.9) converges monotonically to the stabilizing solution of the ARE (2.4) exponentially [20]. It then readily follows that the sequence of time-varying LQR policies (2.8), denoted as $\{K_t\}_{t\in\{N-1,\cdots,0\}}$, converges monotonically to the time-invariant LQR policy K^* as $N \to \infty$. Furthermore, if Q_N satisfies $Q_N > P^*$, then the time-varying LQR policies are stabilizing when treated as frozen. Now, we formally present this convergence result in the following theorem.

Theorem 3.1: Let $A_K^* := A - BK^*$, use $\|\cdot\|_*$ to denote the P^*-induced norm, and define

$$N_0 = \frac{1}{2}\cdot\frac{\log\left(\frac{\|Q_N - P^*\|_*\cdot\kappa_{P^*}\cdot\|A_K^*\|\cdot\|B\|}{\epsilon\cdot\lambda_{\min}(R)}\right)}{\log\left(\frac{1}{\|A_K^*\|_*}\right)} + 1. \tag{3.1}$$

Then, it holds that $\|A_K^*\|_* < 1$ and for all $N \geq N_0$, the control policy K_0^* computed by (2.8) is stabilizing and satisfies $\|K_0^* - K^*\| \leq \epsilon$ for any $\epsilon > 0$.

We provide the proof of Theorem 3.1 in Appendix-B. Theorem 3.1 demonstrates that if selecting $N \sim \mathcal{O}(\log(\epsilon^{-1}))$, then solving the finite-horizon LQR will result in a policy K_0^* that is stabilizing and also ϵ-close to K^*, for any $\epsilon > 0$.

B. Algorithm Design

We propose the RHPG algorithm (cf., Algorithm 1), which first selects N by Theorem 3.1, and then sequentially decomposes the finite-N-horizon LQR problem backward in time. In particular, for every iteration indexed by $h \in \{N-1, \cdots, 0\}$, the RHPG algorithm solves an LQR problem from $t = h$ to $t = N$, where we only optimize for the current policy K_h and fix all the policies $\{K_t\}$ for $t \in \{h+1, \cdots, N-1\}$ to be the convergent solutions generated from earlier iterations. Concretely, for every h, the RHPG algorithm solves the following *quadratic* program in K_h:

$$\begin{aligned} \min_{K_h}\ \mathcal{J}_h(K_h) := \mathbb{E}_{x_h\sim\mathcal{D}}\Bigg[&\sum_{t=h+1}^{N-1} x_t^\top\left(Q + (K_t^*)^\top R K_t^*\right)x_t \\ &+ x_h^\top\left(Q + K_h^\top R K_h\right)x_h + x_N^\top Q_N x_N\Bigg]. \end{aligned} \tag{3.2}$$

Algorithm 1: RHPG

Input: horizon N, max iterations $\{T_h\}$, smoothing radius $\{r_h\}$, stepsizes $\{\eta_h\}$

```
for h = N − 1, ··· , 0 do
    Initialize K_{h,0} arbitrarily (e.g., zero matrix);
    for i = 0, ··· , T_h − 1 do
        // sample PG update via a zeroth-order oracle
        Sample K^+_{h,i} = K_{h,i} + r_h U and
          K^-_{h,i} = K_{h,i} − r_h U, where U is uniformly
          drawn from the surface of a unit sphere, i.e.,
          ||U||_F = 1;
        Sample x_h ~ D and simulate two trajectories
          with policies K^+_{h,i} and K^-_{h,i}, respectively.
          Compute values J_h(K^+_{h,i}) and J_h(K^-_{h,i});
        Compute the estimated PG
          ∇̃J_h(K_{h,i}) = (mn / 2r_h) [J_h(K^+_{h,i}) − J_h(K^-_{h,i})] U
        and update K_{h,i+1} = K_{h,i} − η_h · ∇̃J_h(K_{h,i});
    end
end
Return K_{0,T_0};
```

Due to the quadratic optimization landscape of (3.2) in K_h for every h, applying any PG method with an arbitrary finite initial point (e.g., zero) would lead to convergence to the globally optimal solution of (3.2).

C. Bias of Model-Free Receding-Horizon Control

The RHPG algorithm builds on Bellman's principle of optimality, which requires solving each iteration to the exact optimal solution. However, PG methods can only return an ϵ-accurate solution after a finite number of steps. To generalize Bellman's principle of optimality, we analyze how computational errors accumulate and propagate in the (backward) dynamic programming process. In the theorem below, we show that if one solves every iteration of the RHPG algorithm to the $\mathcal{O}(\epsilon)$-neighborhood of the unique optimum, then the RHPG algorithm will output a policy that is ϵ-close to the infinite-horizon LQR policy K^*.

Theorem 3.2: Choose N according to Theorem 3.1 and assume that one can compute, for all $h \in \{N-1, \cdots, 0\}$ and some $\epsilon > 0$, a policy $\widetilde{K}_h$ that satisfies

$$\left\|\widetilde{K}_h - \widetilde{K}_h^*\right\| \sim \mathcal{O}(\epsilon)\mathcal{O}(1) + \mathcal{O}(\epsilon^{\frac{3}{4}})\mathcal{O}(\texttt{poly}(\text{system parameters})),$$

where $\widetilde{K}_h^*$ is the optimum of the LQR from h to N, after absorbing errors in all previous iterations of Algorithm 1. Then, the RHPG algorithm outputs a control policy $\widetilde{K}_0$ that satisfies $\left\|\widetilde{K}_0 - K^*\right\| \leq \epsilon$. Furthermore, if ϵ is sufficiently small such that $\epsilon < 1 - \|A - BK^*\|_*$, then $\widetilde{K}_0$ is stabilizing.

We illustrate Theorem 3.2 in Figure 1 and defer its proof to Appendix-C. Theorem 3.2 provides specific error tolerance levels for every iteration of the RHPG algorithm to ensure that the output policy is at most ϵ-away from K^*. Then, it remains to establish the sample complexity for the convergence of (zeroth-order) PG methods in every iteration of the algorithm, which is done next.

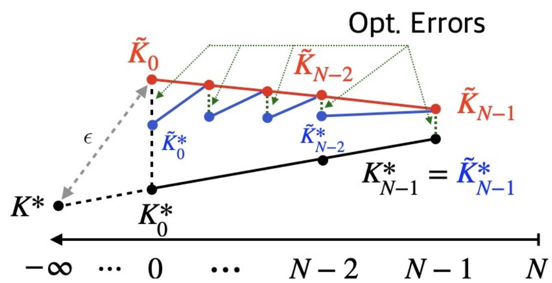

Fig. 1. We show that the output policy $\widetilde{K}_0$ can be made ϵ-close to K^* in two steps. First, Theorem 3.1 proves that K_0^* is ϵ-close to K^* by selecting N accordingly. Then, Theorem 3.2 analyzes the backward propagation of the computational errors from solving each subproblem, denoted as $\delta_t := \widetilde{K}_t - \widetilde{K}_t^*$ for all t, where $\widetilde{K}_t^*$ represents the current optimal LQR policy after absorbing errors from all previous iterations.

D. PG Update and Sample Complexity

We analyze here the sample complexity of the zeroth-order PG update in solving each iteration of the RHPG algorithm. Specifically, the zeroth-order PG update is defined as

$$K_{h,i+1} = K_{h,i} - \eta_h \cdot \widetilde{\nabla}\mathcal{J}_h(K_{h,i}) \tag{3.3}$$

where $\eta_h > 0$ is the stepsize to be determined and $\widetilde{\nabla}\mathcal{J}_h(K_{h,i})$ is the estimated PG sampled from a (two-point) zeroth-order oracle. We formally present the sample complexity result in the following proposition.

Proposition 3.3: For all $h \in \{0, \cdots, N-1\}$, choose a constant smoothing radius $r_h \sim \mathcal{O}(\epsilon)$ and a constant stepsize $\eta_h \sim \mathcal{O}(\epsilon^2)$. Then, the zeroth-order PG update (3.3) converges after $T_h \sim \mathcal{O}(\frac{1}{\epsilon^2}\log(\frac{1}{\delta\epsilon^2}))$ iterations in the sense that $\left\|K_{h,T_h} - \widetilde{K}_h^*\right\| \leq \epsilon$ with a probability of at least $1 - \delta$.

For completeness, we provide a supplementary proof of Proposition 3.3 in Sec. D of [18], which mostly follows existing results in the literature [5]. Combining Theorem 3.2 with Proposition 3.3, we conclude that if we spend $\widetilde{\mathcal{O}}(\epsilon^{-2}\log(\delta^{-1}))$ iterations in solving every subproblem to $\mathcal{O}(\epsilon)$-accuracy with a probability of $1 - \delta$, for all $h \in \{0, \cdots, N-1\}$, then the RHPG algorithm will output a $\widetilde{K}_0$ that satisfies $\|\widetilde{K}_0 - K^*\| \leq \epsilon$ with a probability of at least $1 - N\delta$. By (3.1), this implies that the total iteration complexity of RHPG is also $\widetilde{\mathcal{O}}(\epsilon^{-2}\log(\delta^{-1}))$ with the dependence on various system parameters being polynomial.

We further discuss the tradeoffs in selecting N to balance minimizing the finite-to-infinite error and minimizing errors in inexact dynamic programming in Sec. G of [18]. To compare our sample complexity with the sharpest result in the literature [5], our dependence on ϵ matches that of [5][2]. Both our sample complexity and that of [5] have polynomial dependence on system parameters. However, it is not clear how to compare the polynomial dependencies, and these

[2]Note that the $\widetilde{\mathcal{O}}(\epsilon^{-1})$ sample complexity in [5] is for the convergence in objective value (e.g., $f(K) - f(K^*) \leq \epsilon$), and is equivalent to an $\widetilde{\mathcal{O}}(\epsilon^{-2}) \cdot \mathcal{O}(\texttt{poly}(\text{system parameters}))$ sample complexity for the convergence in policy (i.e., $\|K - K^*\| \leq \epsilon$).

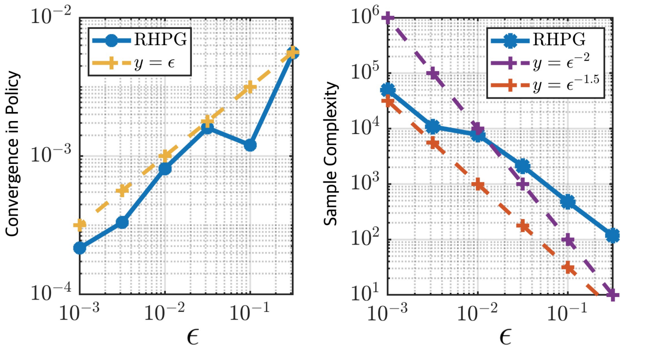

Fig. 2. Under six different ϵ: *Left*: policy error between the output and K^*. *Right*: the total number of calls to the zeroth-order oracle.

polynomial factors might affect the overall computational efficiency of both algorithms in a major way. We leave this comparison as an important topic for future research.

IV. Numerical Experiments

We verify our theories on a scalar linear system studied in [5], where $A = 5$, $B = 0.33$, $Q = R = 1$, and the unique optimal LQR policy is $K^* = 14.5482$. For the PG method in [3], [5] to converge in this simple setting, one has to initialize with a policy K_0 that satisfies $K_0 \in \underline{\mathcal{K}} := \{K \mid 12.12 < K \pm r < 18.18\}$, which is necessary to prevent the zeroth-order oracle with a smoothing radius of r from perturbing K_0 outside of the stabilizing region $\mathcal{K}$ during the first iteration of the PG update. In contrast, we initialize the PG updates in Algorithm 1 with a zero policy $K_h = 0$, set $Q_N = 3$, and choose $N = \texttt{ceil}(\log(\epsilon^{-1}))$ according to (3.1). Furthermore, we choose $r_h = \sqrt{\epsilon}$, select a constant stepsize in each iteration of the RHPG algorithm, and run the algorithm to solve the LQR problem under six different ϵ, namely $\epsilon \in \{10^{-3}, 3.16 \times 10^{-3}, 10^{-2}, 3.16 \times 10^{-2}, 10^{-1}, 3.16 \times 10^{-1}\}$. We apply the zeroth-order PG update in solving every subproblem to $\|\widetilde{K}_h - \widetilde{K}_h^*\| \leq \epsilon$. As shown in Figure 2, the empirical observation of the iteration complexity of RHPG (right) for the convergence in policy (left) is around $\mathcal{O}(\epsilon^{-2})$ under varying ϵ, which corroborates our theoretical findings.

V. Conclusion

We have revisited discrete-time LQR from the perspective of RHPG and provided a fine-grained sample complexity for RHPG to learn a control policy that is both stabilizing and ϵ-close to the optimal LQR policy. Our result demonstrates the potential of RHPG in addressing various tasks in linear control and estimation with streamlined analyses.

References

[1] Alekh Agarwal, Sham M Kakade, Jason D Lee, and Gaurav Mahajan. Optimality and approximation with policy gradient methods in Markov decision processes. In *Conference on Learning Theory*, pages 64–66, 2020.

[2] Kaiqing Zhang, Alec Koppel, Hao Zhu, and Tamer Başar. Global convergence of policy gradient methods to (almost) locally optimal policies. *SIAM Journal on Control and Optimization*, 58(6):3586–3612, 2020.

[3] Maryam Fazel, Rong Ge, Sham M Kakade, and Mehran Mesbahi. Global convergence of policy gradient methods for the linear quadratic regulator. In *International Conference on Machine Learning*, pages 1467–1476, 2018.

[4] Kaiqing Zhang, Bin Hu, and Tamer Başar. Policy optimization for $\mathcal{H}_2$ linear control with $\mathcal{H}_\infty$ robustness guarantee: Implicit regularization and global convergence. *SIAM Journal on Control and Optimization*, 59(6):4081–4109, 2021.

[5] Dhruv Malik, Ashwin Pananjady, Kush Bhatia, Koulik Khamaru, Peter L Bartlett, and Martin J Wainwright. Derivative-free methods for policy optimization: Guarantees for linear quadratic systems. *Journal of Machine Learning Research*, 21(21):1–51, 2020.

[6] Yingying Li, Yujie Tang, Runyu Zhang, and Na Li. Distributed reinforcement learning for decentralized linear quadratic control: A derivative-free policy optimization approach. *IEEE Transactions on Automatic Control*, 67(12):6429–6444, 2021.

[7] Ben Hambly, Renyuan Xu, and Huining Yang. Policy gradient methods for the noisy linear quadratic regulator over a finite horizon. *SIAM Journal on Control and Optimization*, 59(5):3359–3391, 2021.

[8] Kaiqing Zhang, Xiangyuan Zhang, Bin Hu, and Tamer Başar. Derivative-free policy optimization for linear risk-sensitive and robust control design: Implicit regularization and sample complexity. *Advances in Neural Information Processing Systems*, pages 2949–2964, 2021.

[9] Juan C Perdomo, Jack Umenberger, and Max Simchowitz. Stabilizing dynamical systems via policy gradient methods. *Advances in Neural Information Processing Systems*, 34, 2021.

[10] Bin Hu, Kaiqing Zhang, Na Li, Mehran Mesbahi, Maryam Fazel, and Tamer Başar. Toward a theoretical foundation of policy optimization for learning control policies. *Annual Review of Control, Robotics, and Autonomous Systems*, 6(1):123–158, 2023.

[11] Yang Zheng, Yujie Tang, and Na Li. Analysis of the optimization landscape of linear quadratic Gaussian (LQG) control. *Mathematical Programming*, 2023.

[12] Brian DO Anderson and John B Moore. *Optimal Filtering*. Prentice-Hall, 1979.

[13] Brian DO Anderson and John B Moore. *Optimal Control: Linear Quadratic Methods*. Prentice-Hall, Inc., 1990.

[14] Tamer Başar and Pierre Bernhard. *H-infinity Optimal Control and Related Minimax Design Problems: A Dynamic Game Approach*. Birkhäuser, Boston, 1995.

[15] Xiangyuan Zhang, Bin Hu, and Tamer Başar. Learning the Kalman filter with fine-grained sample complexity. In *American Control Conference*, pages 4549–4554, 2023.

[16] Andrew Lamperski. Computing stabilizing linear controllers via policy iteration. In *IEEE Conference on Decision and Control*, pages 1902–1907, 2020.

[17] Feiran Zhao, Xingyun Fu, and Keyou You. Learning stabilizing controllers of linear systems via discount policy gradient. In *Learning for Dynamics and Control*, 2022.

[18] Xiangyuan Zhang and Tamer Başar. Revisiting LQR control from the perspective of receding-horizon policy gradient. *arXiv preprint arXiv:2302.13144*, 2023.

[19] Jingjing Bu, Afshin Mesbahi, Maryam Fazel, and Mehran Mesbahi. LQR through the lens of first order methods: Discrete-time case. *arXiv preprint arXiv:1907.08921*, 2019.

[20] Babak Hassibi, Ali H Sayed, and Thomas Kailath. *Indefinite-Quadratic Estimation and Control: A Unified Approach to H_2 and H_∞ Theories*. SIAM, 1999.

[21] Robert R Bitmead, Michel R Gevers, Ian R Petersen, and R John Kaye. Monotonicity and stabilizability-properties of solutions of the Riccati difference equation: Propositions, lemmas, theorems, fallacious conjectures and counterexamples. *Systems & Control Letters*, 5(5):309–315, 1985.

Appendix

A. Notations

We use $\|X\|$, κ_X, and $\rho(X)$ to denote, respectively, the spectral norm, condition number, and spectral radius of a square matrix X. We define the W-induced norm of X as

$$\|X\|_W^2 := \max_{z \neq 0} \frac{z^\top X^\top W X z}{z^\top W z}.$$

If X is symmetric, we use $X > 0$, $X \geq 0$, $X \leq 0$, and $X < 0$ to denote that X is positive definite, positive semi-definite, negative semi-definite, and negative definite, respectively.

B. Proof of Theorem 3.1

This proof is dual to the proof of Theorem 3.1 in [15]. We first identify one-to-one correspondences between system parameters in LQR and those in Kalman filtering [15]:

LQR:	A	B	Q	R	Q_N
	$\updownarrow$	$\updownarrow$	$\updownarrow$	$\updownarrow$	$\updownarrow$
KF [15]:	$A^\top$	$C^\top$	W	V	X_0

We also identify direct correspondences between our P_t, P^*, K_t, and K^* and [15]'s Σ_{N-t}, Σ^*, L_{N-1-t}, and L^*, respectively. Then, letting

$$\widetilde{P}_t := P_t^* - P^*, \quad \widetilde{R} := R + B^\top P^* B,$$
$$\overline{A} := A - B\widetilde{R}^{-1}B^\top P^* A,$$

and following equations (A.2)-(A.3) of [15], we have

$$\begin{aligned}\widetilde{P}_t &= \overline{A}^\top \widetilde{P}_{t+1}\overline{A} - \overline{A}^\top \widetilde{P}_{t+1}B(\widetilde{R} + B^\top \widetilde{P}_{t+1}B)^{-1}B^\top \widetilde{P}_{t+1}\overline{A} \\ &= \overline{A}^\top \widetilde{P}_{t+1}^{1/2}\big[\boldsymbol{I} + \widetilde{P}_{t+1}^{1/2}B\widetilde{R}^{-1}B^\top \widetilde{P}_{t+1}^{1/2}\big]^{-1}\widetilde{P}_{t+1}^{1/2}\overline{A} \\ &\leq \big[1+\lambda_{\min}(\widetilde{P}_{t+1}^{1/2}B\widetilde{R}^{-1}B^\top \widetilde{P}_{t+1}^{1/2})\big]^{-1}\overline{A}^\top \widetilde{P}_{t+1}\overline{A} \\ &=: \mu_t \overline{A}^\top \widetilde{P}_{t+1}\overline{A}, \end{aligned} \tag{B.1}$$

where $\widetilde{P}_{t+1}^{1/2}$ denotes the unique positive semi-definite (psd) square root of the psd matrix $\widetilde{P}_{t+1}$, $0 < \mu_t \leq 1$ for all t, and $\overline{A}$ satisfies $\rho(\overline{A}) < 1$. We now use $\|\cdot\|_*$ to represent the P^*-induced matrix norm and invoke Theorem 14.4.1 of [20], where our $\widetilde{P}_t$, $\overline{A}^\top$ and P^* correspond to $P_i - P^*$, F_p and W in [20], respectively. By Theorem 14.4.1 of [20] and (B.1), we obtain $\|\overline{A}\|_* < 1$ and given that $\mu_t \leq 1$,

$$\|\widetilde{P}_t\|_* \leq \|\overline{A}\|_*^2 \cdot \|\widetilde{P}_{t+1}\|_*.$$

Therefore, the convergence is exponential such that $\|\widetilde{P}_t\|_* \leq \|\overline{A}\|_*^{2(N-t)} \cdot \|\widetilde{P}_N\|_*$. As a result, the convergence of $\widetilde{P}_t$ to $\mathbf{0}$ in spectral norm can be characterized as

$$\|\widetilde{P}_t\| \leq \kappa_{P^*} \cdot \|\widetilde{P}_t\|_* \leq \kappa_{P^*} \cdot \|\overline{A}\|_*^{2(N-t)} \cdot \|\widetilde{P}_N\|_*,$$

where we have used κ_X to denote the condition number of X. That is, to ensure $\|\widetilde{P}_1\| \leq \epsilon$, it suffices to require

$$N \geq \frac{1}{2} \cdot \frac{\log\big(\frac{\|\widetilde{P}_N\|_* \cdot \kappa_{P^*}}{\epsilon}\big)}{\log\big(\frac{1}{\|\overline{A}\|_*}\big)} + 1. \tag{B.2}$$

Furthermore, since (A, B) is stabilizable, $(A, Q^{1/2})$ is detectable, and $Q_N > P^*$, the closed-loop system at any time $t \in \{0, \cdots, N-1\}$ is exponentially asymptotically stable such that the time-invariant (frozen) LQR policy satisfies [21]

$$\rho(A - B(R + B^\top P_{t+1}^* B)^{-1}B^\top P_{t+1}^* A) < 1.$$

Lastly, we show that the (monotonic) convergence of the LQR gain to the time-invariant LQR gain follows from the convergence of P_t^* to P^*. Similar to (A.5) of [15], this can be verified through:

$$\begin{aligned} K_t^* - K^* &= (R + B^\top P_{t+1}^* B)^{-1}B^\top P_{t+1}^* A \\ &\quad - (R + B^\top P^* B)^{-1}B^\top P^* A \\ &= \big[(R + B^\top P_{t+1}^* B)^{-1} - (R + B^\top P^* B)^{-1}\big]B^\top P^* A \\ &\quad + (R + B^\top P_{t+1}^* B)^{-1}B^\top (P_{t+1}^* - P^*)A \\ &= (R + B^\top P_{t+1}^* B)^{-1}B^\top (P^* - P_{t+1}^*)BK^* \\ &\quad - (R + B^\top P_{t+1}^* B)^{-1}B^\top (P^* - P_{t+1}^*)A \\ &= (R + B^\top P_{t+1}^* B)^{-1}B^\top (P^* - P_{t+1}^*)(BK^* - A) \end{aligned} \tag{B.3}$$

Hence, we have $\|K_t^* - K^*\| \leq \frac{\|\overline{A}\| \cdot \|B\|}{\lambda_{\min}(R)} \cdot \|P_{t+1}^* - P^*\|$ and

$$\|K_0^* - K^*\| \leq \frac{\|\overline{A}\| \cdot \|B\|}{\lambda_{\min}(R)} \cdot \|\widetilde{P}_1\|.$$

Substituting ϵ in (B.2) with $\frac{\epsilon \cdot \lambda_{\min}(R)}{\|\overline{A}\| \cdot \|B\|}$ completes the proof.

C. Proof of Theorem 3.2

This proof is dual to the proof of Theorem 3.3 in [15]. First, according to Theorem 3.1, we select

$$N = \frac{1}{2} \cdot \frac{\log\big(\frac{2\|Q_N - P^*\|_* \cdot \kappa_{P^*} \cdot \|A_K^*\| \cdot \|B\|}{\epsilon \cdot \lambda_{\min}(R)}\big)}{\log\big(\frac{1}{\|A_K^*\|_*}\big)} + 1, \tag{C.4}$$

where $A_K^* := A - BK^*$. This ensures that K_0^* is stabilizing and $\|K_0^* - K^*\| \leq \epsilon/2$. Then, it remains to show that the output $\widetilde{K}_0$ satisfies $\|\widetilde{K}_0 - K_0^*\| \leq \epsilon/2$.

Recall that the RDE (2.9) is a backward iteration starting with $P_N^* = Q_N > 0$, and can also be represented as:

$$P_t^* = A^\top P_{t+1}^*(A - BK_t^*) + Q \tag{C.5}$$
$$= (A - BK_t^*)^\top P_{t+1}^*(A - BK_t^*) + (K_t^*)^\top RK_t^* + Q. \tag{C.6}$$

Moreover, for any K_t, we define the Lyapunov equation:

$$P_t = (A - BK_t)^\top P_{t+1}(A - BK_t) + K_t^\top RK_t + Q. \tag{C.7}$$

Furthermore, for clarity of the proof, we define/recall:

- K_t^*: Exact LQR policy at time t defined in (2.8)
- $\widetilde{K}_t^*$: Optimal policy of the current cost-to-go function, absorbing errors in all prior steps
- $\widetilde{K}_t$: An approximation of $\widetilde{K}_t^*$ obtained by the PG update (3.3)
- $\delta_t := \widetilde{K}_t - \widetilde{K}_t^*$: Policy optimization error at time t
- $\widetilde{P}_t^*$: Generated by (C.6) with $K_t^* = \widetilde{K}_t^*$ and $P_{t+1}^* = \widetilde{P}_{t+1}$.

We argue that $\|\widetilde{K}_0 - K_0^*\| \leq \epsilon/2$ can be achieved by carefully controlling δ_t for all t. At $t = 0$, it holds that

$$\|\widetilde{K}_0 - K_0^*\| \leq \|\widetilde{K}_0^* - K_0^*\| + \|\delta_0\|,$$

where substituting K_t^* and K^* in (B.3), respectively, with $\widetilde{K}_0^*$ and K_0^* leads to

$$\widetilde{K}_0^* - K_0^* = (R + B^\top \widetilde{P}_1 B)^{-1}B^\top (P_1^* - \widetilde{P}_1)(BK_0^* - A).$$

By Theorem 3.1, K_0^* is stabilizing and it holds that $P_1^* \geq P^*$ due to $Q_N > P^*$. Next, we require $\|P_1^* - \widetilde{P}_1\| \leq \|P^*\|$ to ensure the positive definiteness of $\widetilde{P}_1$ and derive

$$\|\widetilde{K}_0^* - K_0^*\| \leq \frac{\|A - BK_0^*\| \cdot \|B\|}{\lambda_{\min}(R)} \cdot \|P_1^* - \widetilde{P}_1\|. \quad \text{(C.8)}$$

Define the helper constant

$$C_1 := \frac{\varphi \cdot \|B\|}{\lambda_{\min}(R)} > 0, \ \varphi := \max_{t \in \{0, \cdots, N-1\}} \|A - BK_t^*\|.$$

Next, require $\|\delta_0\| \leq \epsilon/4$ and $\|\widetilde{K}_0^* - K_0^*\| \leq \epsilon/4$ to fulfill $\|\widetilde{K}_0 - K_0^*\| \leq \epsilon/2$. By (C.8), it suffices to require

$$\|P_1^* - \widetilde{P}_1\| \leq \min\left\{\|P^*\|, \frac{\epsilon}{4C_1}\right\}. \quad \text{(C.9)}$$

Subsequently, by (C.7), we have

$$P_1^* - \widetilde{P}_1 = (P_1^* - \widetilde{P}_1^*) + (\widetilde{P}_1^* - \widetilde{P}_1). \quad \text{(C.10)}$$

The first difference term on the RHS of (C.10) is

$$\begin{aligned} &P_1^* - \widetilde{P}_1^* = A^\top P_2^*(A - BK_1^*) - A^\top \widetilde{P}_2(A - B\widetilde{K}_1^*) \\ &= A^\top (P_2^* - \widetilde{P}_2)(A - BK_1^*) + A^\top \widetilde{P}_2 B(\widetilde{K}_1^* - K_1^*). \quad \text{(C.11)} \end{aligned}$$

Moreover, the second term on the RHS of (C.10) is

$$\begin{aligned} \widetilde{P}_1^* - \widetilde{P}_1 &= (A - B\widetilde{K}_1^*)^\top \widetilde{P}_2(A - B\widetilde{K}_1^*) + (\widetilde{K}_1^*)^\top R\widetilde{K}_1^* \\ &\quad - (A - B\widetilde{K}_1)^\top \widetilde{P}_2(A - B\widetilde{K}_1) - (\widetilde{K}_1)^\top R\widetilde{K}_1 \\ &= -(\widetilde{K}_1^*)^\top B^\top \widetilde{P}_2 A - A^\top \widetilde{P}_2 B\widetilde{K}_1^* + (\widetilde{K}_1^*)^\top (R + B^\top \widetilde{P}_2 B)\widetilde{K}_1^* \\ &\quad + \widetilde{K}_1^\top B^\top \widetilde{P}_2 A + A^\top \widetilde{P}_2 B\widetilde{K}_1 - \widetilde{K}_1^\top (R + B^\top \widetilde{P}_2 B)\widetilde{K}_1 \\ &= \big[(R + B^\top \widetilde{P}_2 B)^{-1} B^\top \widetilde{P}_2 A - \widetilde{K}_1^*\big]^\top (R + B^\top \widetilde{P}_2 B) \cdot \\ &\quad \big[(R + B^\top \widetilde{P}_2 B)^{-1} B^\top \widetilde{P}_2 A - \widetilde{K}_1^*\big] \\ &\quad - \big[(R + B^\top \widetilde{P}_2 B)^{-1} B^\top \widetilde{P}_2 A - \widetilde{K}_1\big]^\top (R + B^\top \widetilde{P}_2 B) \cdot \\ &\quad \big[(R + B^\top \widetilde{P}_2 B)^{-1} B^\top \widetilde{P}_2 A - \widetilde{K}_1\big] \quad \text{(C.12)} \\ &= -\delta_1^\top (R + B^\top \widetilde{P}_2 B)\delta_1, \quad \text{(C.13)} \end{aligned}$$

where (C.12) follows from completion of squares. Thus, combining (C.8), (C.10), (C.11), and (C.13) yields

$$\begin{aligned} &\|P_1^* - \widetilde{P}_1\| \\ &\leq \|P_2^* - \widetilde{P}_2\| \cdot \varphi\|A\| + \|\widetilde{K}_1^* - K_1^*\| \cdot \|B\| \cdot \|\widetilde{P}_2\| \cdot \|A\| \\ &\quad + \|\delta_1\|^2 \|R + B^\top \widetilde{P}_2 B\| \\ &\leq \|A\| \cdot \big[\varphi + C_1 \cdot \|B\| \cdot \|\widetilde{P}_2\|\big] \cdot \|P_2^* - \widetilde{P}_2\| \\ &\quad + \|\delta_1\|^2 \|R + B^\top \widetilde{P}_2 B\|. \quad \text{(C.14)} \end{aligned}$$

Now, we require

$$\|P_2^* - \widetilde{P}_2\| \leq \min\left\{\|P^*\|, \frac{\|P^*\|}{C_2}, \frac{\epsilon}{4C_1C_2} \cdot \right\} \quad \text{(C.15)}$$

$$\|\delta_1\| \leq \min\left\{\sqrt{\frac{\|P^*\|}{C_3}}, \frac{1}{2}\sqrt{\frac{\epsilon}{C_1C_3}}\right\}, \quad \text{(C.16)}$$

where C_2 and C_3 are positive constants defined as

$$\begin{aligned} C_2 &:= 2\|A\| \cdot \big[\varphi + C_1 \cdot \|B\| \cdot \big(\|Q_N\| + \|P^*\|\big)\big] > 0 \\ C_3 &:= 2\big[\|R\| + \|B\|^2 \big(\|Q_N\| + \|P^*\|\big)\big] > 0. \end{aligned}$$

Then, conditions (C.15) and (C.16) are sufficient for (C.9) (and thus for $\|\widetilde{K}_0 - K_0^*\| \leq \epsilon/2$) to hold. Subsequently, we can propagate the required accuracies in (C.15) and (C.16) forward in time. Specifically, we iteratively apply the arguments in (C.14) (i.e., by plugging quantities with subscript t into the LHS of (C.14) and plugging quantities with subscript $t+1$ into the RHS of (C.14)) to obtain the result that if at all $t \in \{2, \cdots, N-1\}$, we require

$$\|P_t^* - \widetilde{P}_t\| \leq \min\left\{\|P^*\|, \frac{\|P^*\|}{C_2^{t-1}}, \frac{\epsilon}{4C_1C_2^{t-1}}\right\} \quad \text{(C.17)}$$

$$\|\delta_t\| \leq \min\left\{\sqrt{\frac{\|P^*\|}{C_3}}, \sqrt{\frac{\|P^*\|}{C_2^{t-2}C_3}}, \frac{1}{2}\sqrt{\frac{\epsilon}{C_1C_2^{t-2}C_3}}\right\},$$

then (C.15) holds true and therefore (C.9) is satisfied.

We now compute the required accuracy for δ_{N-1}. Note that $P_{N-1}^* = \widetilde{P}_{N-1}^*$ since no prior computational errors happened at $t = N$. By (C.14), the distance between P_{N-1}^* and $\widetilde{P}_{N-1}$ can be bounded as

$$\|P_{N-1}^* - \widetilde{P}_{N-1}\| = \|\widetilde{P}_{N-1}^* - \widetilde{P}_{N-1}\| \leq \|\delta_{N-1}\|^2 \cdot C_3.$$

To fulfill the requirement (C.17) for $t = N-1$, which is

$$\|P_{N-1}^* - \widetilde{P}_{N-1}\| \leq \min\left\{\|P^*\|, \frac{\|P^*\|}{C_2^{N-2}}, \frac{\epsilon}{4C_1C_2^{N-2}}\right\},$$

it suffices to let

$$\|\delta_{N-1}\| \leq \min\left\{\sqrt{\frac{\|P^*\|}{C_3}}, \sqrt{\frac{\|P^*\|}{C_2^{N-2}C_3}}, \frac{1}{2}\sqrt{\frac{\epsilon}{C_1C_2^{N-2}C_3}}\right\}. \quad \text{(C.18)}$$

Finally, we analyze the worst-case complexity of RHPG by computing, at the most stringent case, the required size of $\|\delta_t\|$. When $C_2 \leq 1$, the most stringent dependence of $\|\delta_t\|$ on ϵ happens at $t = 0$, which is of the order $\mathcal{O}(\epsilon)$, and the dependences on system parameters are $\mathcal{O}(1)$. We then analyze the case where $C_2 > 1$, where the requirement on $\|\delta_0\|$ is still $\mathcal{O}(\epsilon) \cdot \mathcal{O}(1)$. Note that in this case, $\|\delta_{N-1}\| \leq \|\delta_t\|$ for all $t \in \{1, \cdots, N-1\}$ and by (C.18):

$$\|\delta_{N-1}\| \sim \mathcal{O}\Big(\sqrt{\frac{\epsilon}{C_1C_2^{N-2}C_3}}\Big). \quad \text{(C.19)}$$

Since we require N to satisfy (C.4), the dependence of $\|\delta_{N-1}\|$ on ϵ in (C.19) becomes $\|\delta_{N-1}\| \sim \mathcal{O}(\epsilon^{\frac{3}{4}})$ with additional polynomial dependences on system parameters, but one can observe that the dependence on ϵ is still milder than the requirement for $\|\delta_0\|$. Therefore, it suffices to require error bound for all t to be

$$\|\delta_t\| \sim \mathcal{O}(\epsilon) \cdot \mathcal{O}(1) + \mathcal{O}(\epsilon^{\frac{3}{4}}) \cdot \mathcal{O}(\texttt{poly}(\text{system parameters}))$$

to reach the ϵ-neighborhood of the infinite-horizon LQR policy. Lastly, for $\widetilde{K}_0$ to be stabilizing, it suffices to select a sufficiently small ϵ to satisfy the crude bound of

$$\epsilon < 1 - \|A - BK^*\|_* \Longrightarrow \|A - B\widetilde{K}_0\|_* < 1.$$

This completes the proof.